Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren

Bubenhofer, Noah ; Scharloth, Joachim

Preprint von: Bubenhofer, Noah/Scharloth, Joachim (2013): Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren. In: Warnke, Ingo H./Meinhof, Ulrike/Reisigl, Martin (Hrsg.): Diskurslinguistik im Spannungsfeld von Deskription und Kritik. Berlin: Akademie-Verlag. S. 147-168.

Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren

Noah Bubenhofer/Joachim Scharloth (Dresden)

1 Warum Korpuslinguistik?

Neuere diskurslinguistische Arbeiten kommen selten umhin, korpuslinguistische Methoden anzuwenden, um ihre Datengrundlage zu analysieren, zumal wenn die Datengrundlage aus geschriebenen Texten besteht und in umfangreicher Menge in digitaler Form zur Verfügung steht. Bei den eingesetzten Methoden handelt es sich einerseits um mehr oder weniger elaborierte Suchstrategien, um Belege für bestimmte vermutete Phänomene zu finden (Wortverwendungen, Phrasen, Topoi, Metaphern etc.) und deren Distribution im Korpus zu erfassen, andererseits aber auch um Methoden, um überhaupt erst Hypothesen über die Existenz von Phänomenen auf der Basis linguistischer Analysen zu gewinnen. Mit Hilfe dieser Methoden kann eruiert werden, welche Lexeme oder syntagmatischen Muster typisch für bestimmte Bereiche des Korpus oder welche Texte bezüglich sprachlicher Merkmale ähnlich oder unähnlich sind. Entsprechende Befunde können dann diskurslinguistisch gedeutet werden und dienen der empirischen Unterfütterung diskurslinguistischer Hypothesen. Ob dabei die Deutung eher dem deskriptiven oder dem kritischen Paradigma folgt, ist dabei vorerst zweitrangig. Wir wollen aber zeigen, dass eine quantitativ-empirische Unterfütterung sowohl deskriptiven als auch kritischen Ansätzen hilft, ein besseres Argument für die Interpretation zu haben.

So anerkannt der Einsatz korpuslinguistischer Methoden in der Diskurslinguistik zu sein scheint, so werden jedoch auch immer wieder Grenzen der Korpuslinguistik genannt:

1. Ein Diskurs wird nicht zwingend durch ein Korpus mit einer Anzahl von Texten repräsentiert. Diskurse sind vielmehr als Aussagensysteme zu begreifen, die quer zu Texten liegen (vgl. Spitzmüller/Warnke 2011: 25, 88, 91).
2. Diskurse sind multimodal und nicht auf schriftliche Texte beschränkt. Korpuslinguistische Verfahren weisen deshalb einen blinden Fleck auf, wenn sie andere Medien außer Acht lassen (vgl. Spitzmüller/Warnke 2011: S. 39).
3. Korpuslinguistische Methoden arbeiten frequenzorientiert. Doch nicht alles, was diskurslinguistisch relevant ist, ist in einem Korpus frequent. Deshalb müssen quantitative Ansätze durch qualitative Analysemethoden ergänzt werden (vgl. Spitzmüller/Warnke 2011: 39).

Auf die ersten beiden Einwände wollen wir in diesem Aufsatz nur kurz eingehen: Dass die Modellierung eines Diskurses über ein abgeschlossenes, thematisch definiertes Textkorpus zu kurz greifen könnte, ist sehr bedenkenswert, und das Postulat, Diskurse eher als regelmäßige Art des Aussagens aufzufassen, eine wichtige Anregung. Diese spricht jedoch keineswegs gegen korpuslinguistische Verfahren, sondern bedingt andere, dynamischere Korpusdefinitionen und Analysemethoden. Uns scheint gerade die klassischste Darstellungsform der Korpuslinguistik von Textbelegen als Key-Word-in Context-Listen eine solche Lesart zu begünstigen. Gerade diese Darstellungsform durchbricht die Einheit des Textes und lässt Regelmäßigkeiten über Texte hinweg sichtbar werden (vgl. dazu ausführlich Bubenhofer 2009: 32 ff.).

Auch das Desiderat, Diskurslinguistik multimodal aufzufassen, ist plausibel. Die Korpuslinguistik auf die Analyse von geschriebenen Texten zu reduzieren ist jedoch falsch: Es sind momentan noch technische und rechtliche Hürden vorhanden, auch gesprochene Sprache in großem Umfang korpuslinguistisch aufzubereiten, diese Hürden sind jedoch nicht unüberwindbar. Zudem lassen sich textorientierte Analysen jederzeit in empirisch-quantitativer Weise mit anderen Daten, z.B. aus Bild- oder Videoanalysen, ergänzen. Da die Korpuslinguistik einem empirisch-quantitativen Paradigma verpflichtet ist, ist sie an die breite Palette weiterer quantitativer Analyseverfahren anschließbar. So spricht nichts dagegen, die Ergebnisse einer Inhaltsanalyse von Bildcodes mit korpuslinguistischen Analysen zu vergleichen.[1]

Im Folgenden möchten wir vor allem auf den letzten Einwand eingehen: die Frage nach dem Nutzen und den Möglichkeiten empirisch-quantitativer Analysemethoden der Korpuslinguistik. Wir sehen korpuslinguistisches Denken als eigenständigen Denkstil, der die Aspekte der Empirie, Quantität und Textorientierung miteinander verbindet. Diese drei Aspekte sind als methodische Folge eines „korpuspragmatischen Paradigmas" zu verstehen. Wir haben bereits dargelegt (Scharloth/Bubenhofer 2011), dass wir die Forschung als korpuspragmatische verstehen, die wie folgt charakterisiert ist: Die Korpuspragmatik deutet signifikant häufig auftretende sprachliche Muster in Korpora als Ergebnis rekurrenter Sprachhandlungen der Autorinnen und Autoren der im Korpus enthaltenen Texte bzw. der sie autorisierenden Institutionen und Gruppen. Sie geht davon aus, dass sich pragmatische Informationen „im pragmatischen Mehrwert oder Gebrauchswert von Einheiten aller sprachlicher Strukturbereiche" (Feilke 2000: 78) zeichenhaft manifestiert. Damit werden pragmatische Spuren an der sprachlichen Oberfläche als Muster, in die sich ein Gebrauchswert eingeschrieben hat, sichtbar (vgl. zu dieser Rehabilitierung der Textoberfläche auch den Sammelband von Feilke/Linke 2009). Diese „Sprachgebrauchsmuster" (Bubenhofer 2009) werden damit als Ergebnis von sprachlich-sozialem Handeln gelesen und gedeutet.

2 Prämissen einer korpuslinguistischen Diskursanalyse

2.1 Datengeleitete Ansätze

Attraktiv an der Orientierung an sprachlichen Mustern auf der Textoberfläche sind die methodischen Möglichkeiten, die sich dadurch ergeben: Dank der Verfügbarkeit von großen Korpora und der maschinellen Verarbeitung von Textdaten wird es möglich, quantitativ zu arbeiten und Algorithmen zu entwickeln, um die Musterhaftigkeit in den Daten induktiv zu entdecken. Digitale Korpora sind somit nicht nur „Belegsammlungen oder Zettelkästen in elektronischer Form", sondern ermöglichen eine eigene „korpuslinguistische Perspektive" (Perkuhn/Belica 2006: 2).

Was ist mit dieser „korpuslinguistischen Perspektive" gemeint? Die oben gemachten Ausführungen deuten es bereits an. Offensichtlich ist das quantitative Vorgehen: Nicht der Einzelbeleg interessiert, sondern die Musterhaftigkeit von Belegen, die als Ergebnis rekurrenter Sprachhandlungen gelesen werden können. Doch es kommt ein weiterer

[1] Obwohl sowohl die Inhaltsanalyse als auch die hier dargestellte Korpuslinguistik einem empirisch-quantitativen Paradigma verpflichtet sind, ist eine Kombination solcher Methoden keinesfalls trivial, da sich die Methoden im Detail stark unterscheiden können. Trotzdem: Auch methodisch unterschiedlich erhobene Ergebnisse können, so lange sie quantitativ sind und aus empirischer Beobachtung stammen, einfacher miteinander verglichen werden als beispielsweise quantitativ-statistisch und qualitativ-hermeneutisch erhobene Ergebnisse.

Aspekt hinzu: In Ergänzung zu den klassischen datenbasierten korpuslinguistischen Analysen, die der Überprüfung von Forschungshypothesen dienen („corpus-based"-Paradigma), ist es fruchtbar, ein induktives Vorgehen bei der datengeleiteten Analyse („corpus-driven"-Paradigma) hinzuzuziehen. Dieses Paradigma wird von Tognini-Bonelli (2001:

84ff.) vor dem Hintergrund der Arbeiten von Sinclair (1991) expliziert und im deutschen Sprachraum in mehreren Arbeiten (z.B. Belica/Steyer 2006; Steyer 2004; Bubenhofer 2009) verbreitet. Statt eine Hypothese mit vorher festgelegten Analysekategorien zu überprüfen, werden in einem Korpus sämtliche Zeichenkonfigurationen berechnet, die sich bei der Anwendung vorher festgelegter Algorithmen ergeben. Beispielsweise werden alle unterschiedlichen Kovorkommen von Wörtern berechnet. Im Vergleich mit einem Referenzkorpus können nun die Kovorkommen berechnet werden, die im Untersuchungskorpus besonders auffällig – also statistisch gesehen überzufällig häufig – sind. Diese Muster werden im Anschluss kategorisiert. Damit geraten häufig Evidenzen in den Fokus, die entweder quer zu den vorher existierenden Erwartungen stehen und die Grundlage für neue Hypothesen sind, oder im besten Fall sogar solche Evidenzen, die die Bildung neuer interpretativer linguistischer Analysekategorien nahelegen. Gleichwohl ist auch das datengeleitete Paradigma keinesfalls vollständig induktiv und frei von Ausgangshypothesen: So lautet die Ausgangshypothese beispielsweise, dass häufige Zeichenkonfigurationen für die Untersuchungsfrage bedeutend sind. Jedoch gehen datengeleitete Verfahren von sehr viel offeneren Hypothesen aus als hypothesengeleitete Verfahren.

Methodisch gesehen bietet sich für induktive Analysen die Berechnung von n-Grammen an: N-Gramme sind Einheiten einer durch *n* bestimmten Anzahl aufeinander folgender Wörter (Manning/Schütze 2002: 192 ff.). Durch kombinatorische Zählverfahren lassen sich in einem Korpus leicht alle vorhandenen n-Gramme berechnen. Zudem können, wenn das Korpus vorgängig entsprechend annotiert wurde, neben den Wortformen die n-Gramme auch aus Wortart-Informationen oder einer Kombination von Wortform und Wortart bestehen. Solche „komplexen n-Gramme" (Scharloth/Bubenhofer 2011; Bubenhofer u. a. 2009) sind abstrakter als reine Wortformen-n-Gramme, da sie unterschiedliche Varianten des gleichen zugrundeliegenden morphosyntaktischen Musters zusammenfassen. Weiter unten werden wir im Rahmen von Beispielanalysen dieses Verfahren illustrieren.

2.2 Empirie und Quantität

Während eine empirische Fundierung in der diskurslinguistischen Methodendiskussion kaum bestritten wird, ist die Frage nach dem Gleichgewicht zwischen quantitativen und qualitativen Vorgehensweisen diskussionswürdig. Die Korpuslinguistik ist per se empirisch ausgerichtet (Lemnitzer/Zinsmeister 2006: 15), doch ist sie nicht zwingend quantitativ. Bei „korpusgestützten Ansätzen" (Lemnitzer/Zinsmeister 2006: 37) dienen Korpusdaten höchstens als zusätzliche Quelle für Evidenz, um bestehende Hypothesen zu prüfen oder Belege dafür zu finden.

Man würde jedoch kapitale Chancen der Datenanalyse verpassen, wenn man auf quantitative Methoden verzichten würde. Und dies hängt unmittelbar mit der Empirie zusammen: Der Reiz an empirischem Arbeiten liegt darin, Tendenzen in realen Daten festzustellen, die erst wahrnehmbar sind, wenn man einen Schritt zurück tritt und das gesamte Bild sieht – allerdings verbunden mit der Gewissheit, dass doch jedes Einzelvorkommen etwas zu diesem Bild beiträgt. Das Bild fasst die Daten derart

zusammen, dass sie in ihrer Gesamtheit erfasst werden können, ohne – und das ist der Gewinn korpuslinguistischer Analyseverfahren – zu grob zu vereinfachen.

Dazu ein Beispiel: Mit statistischen Mitteln sind die typischen syntagmatischen Muster berechenbar, die ein beliebiges Lemma in einem Korpus aufweist. Gemäß der Kookkurrenzdatenbank „CCDB“ (Belica 2001), deren Datengrundlage das Deutsche Referenzkorpus ist (Institut für Deutsche Sprache 2010), sind das für *Maßnahme* die folgenden Muster (nur die ersten sechs Knoten dargestellt):

L	*R*	*LLR*	*Kollokatoren*	*#*		*syntagmatisches Muster*
1	5	19420	ergreifen notwendigen	101	92%	alle|die notwendigen [...] Maßnahmen [zu] ergreifen um
			ergreifen nötigen	29	86%	die|alle nötigen [...] Maßnahmen [zu] ergreifen um die
			ergreifen angemessene	26	92%	und angemessene [...] Maßnahmen [zu] ergreifen damit
			ergreifen	1952	84%	Maßnahmen [zu] ergreifen um
1	5	11561	ergriffen werden würden müßten	1	100%	würden müßten Maßnahmen ... ergriffen werden
			ergriffen werden würden	6	50%	werden ... Maßnahmen [...] ergriffen würden
			ergriffen werden müßten	26	69%	müßten [auch ...] Maßnahmen [...] ergriffen werden um die
			ergriffen werden	485	68%	Maßnahmen [...] ergriffen [...] werden
			ergriffen würden	86	48%	nicht ... Maßnahmen [...] ergriffen [...] würden
			ergriffen müßten	32	78%	müßten [...] Maßnahmen [...] ergriffen werden um
			Ergriffen	1269	80%	Maßnahmen [...] ergriffen werden
-1	-1	10629	vertrauensbildende Als	24	83%	Als [...] vertrauensbildende Maßnahme
			vertrauensbildende setzen als	1	100%	vertrauensbildende Maßnahmen setzen als
			vertrauensbildende setzen	16	100%	und vertrauensbildende Maßnahmen [zu] setzen
			vertrauensbildende als	155	81%	als [...] vertrauensbildende Maßnahme
			Vertrauensbildende	646	56%	vertrauensbildende [...] Maßnahmen
-1	-1	8920	konkrete erste	36	47%	Als|als erste [...] konkrete [...] Maßnahme ... die
			konkrete Arbeitslosigkeit	17	88%	auf konkrete [...] Maßnahmen zur|zum|gegen Abbau|die|der Arbeitslosigkeit
			konkrete umzusetzen	11	100%	in konkrete [...] Maßnahmen [zur ...] umzusetzen
			Konkrete	1048	78%	konkrete [...] Maßnahmen
-1	-1	8666	solche nötig sei nicht	1	100%	solche Massnahme nicht nötig sei
			solche nötig sei	3	33%	solche Massnahme ... nötig sei
			solche nötig nicht	8	50%	eine solche Massnahme [...] nicht [mehr] nötig
			solche nötig	18	44%	solche Maßnahmen [... es|nicht] nötig
			solche sei nicht	19	26%	Eine solche Maßnahme sei [...] nicht
			solche sei	69	39%	Eine|eine solche [...] Maßnahme [...] sei
			solche nicht	231	41%	solche [...] Maßnahmen [...] nicht
			Solche	1540	45%	solche [...] Maßnahmen
-1	-1	8593	flankierenden Paket	4	100%	Paket der flankierenden Massnahmen
			flankierenden bilateralen	23	60%	bei die|der|den flankierenden Massnahmen zu|zum den bilateralen Verträgen ...
			Flankierenden	550	71%	die|den flankierenden [...] Massnahmen

Die Tabelle stellt nur einen Ausschnitt dar, enthält aber eine riesige Anzahl an Informationen: In der Spalte „Kollokatoren“ sind fett gedruckt die statistisch signifikantesten Wörter aufgelistet, die zusammen mit *Maßnahme* vorkommen. So ist die auffälligste Wortform *ergreifen* (vor *ergriffen*). Dazu gibt es sekundäre Kollokatoren: *angemessene, nötigen, notwendigen*. In der Spalte # sind die absoluten Frequenzen dieser Verbindungen ersichtlich: *Maßnahmen ergreifen* ist in diesem Cluster die häufigste Kollokation. In den Spalten „L“ und „R“ wird angeben, in welchem Bereich links und rechts des Suchwortes sich der Kollokator *ergreifen* bewegt: Von einem Wort davor bis fünf Wörter danach reicht das Fenster. Weitere Informationen zur Verwendungsweise sind über die „syntagmatischen Muster“ ersichtlich: In 84% der 1952 Fälle von *Maßnahme + ergreifen* wird das Muster *Maßnahmen [zu] ergreifen um*

realisiert. Etwas variabler ist die Verwendung der Kollokation *Maßnahme + ergreifen + nötigen*: Es lautet *die|alle nötigen [...] Maßnahmen [zu] ergreifen um die*. Das Syntagma wird in 86% der Fälle eingeführt mit *die nötigen* oder *alle nötigen*, worauf beliebiger Text folgt und dann die Wortgruppe *Maßnahmen [zu] ergreifen um die*.

Die Funktionsweise statistischer Signifikanzmaße zeigt sich anschaulich beim Cluster *Maßnahme + vertrauensbildende*. Obwohl die absolute Häufigkeit dieser Kollokation mit 646 niedriger ist als die der nachfolgenden Cluster, ist der statistische Log-Likelihood-Wert[2] höher, da die Eigenfrequenz von *vertrauensbildende* im Korpus in die Berechnung mit einbezogen wurde: Die ist weniger hoch als beispielsweise von *ergreifen*, doch trotzdem ist die Kombination mit *Maßnahme* häufiger, als es eine zufällige Verteilung voraussagen würde.

Am Kookkurrenzprofil lässt sich der Sprachgebrauch von *Maßnahme* genau in der oben postulierten Weise beobachten: Die quantitative Analyse fasst die Daten derart zusammen, dass es möglich wird, sie in ihrer Gesamtheit erfassen zu können, ohne Details zu vernachlässigen: Der statistische Algorithmus liefert nicht einen alles einebnenden „Durchschnitt" der Daten, sondern ein differenziertes Bild, das erst durch die quantitative Analyse möglich wird. Die qualitative und detaillierte Analyse einer Handvoll Belege, auch wenn es 1000 sind, kann dieses Bild nicht liefern.

Welche diskurslinguistisch interessanten Aspekte sich alleine in diesem Kookkurrenzprofil verbergen, muss wohl nur angedeutet werden: Es zeigt sehr genau auf, welche Strategien der diskursiven Verschleierung unter Vortäuschung von Aktivismus möglich sind. Alleine, dass *Maßnahme* so musterhaft verwendet wird, zeigt diese pragmatische Funktion; alternative Formulierungsmuster äußern sich wahrscheinlich weniger formelhaft und gehören deshalb weniger zum Standardrepertoire von z.B. politischen Akteuren.

Interessant ist sodann der Vergleich dieses Profils mit anderen Daten wie beispielsweise dem *GerMov-Korpus*, einem Korpus zur gesprochenen und geschriebenen Sprache der 68er-Bewegung. Das Korpus wurde im Rahmen einer umfangreichen Studie zum Einfluss von 68er-Bewegung und Alternativmilieu auf die Kommunikationsgeschichte der Bundesrepublik Deutschland erstellt (Scharloth 2011) und gliedert sich in zwei Subkorpora: ein Subkorpus, das ausschließlich Tonbandprotokolle der 68er-Bewegung enthält, und ein Subkorpus mit Flugblättern. Das Korpus der Tonbandprotokolle enthält 29 Protokolle mit 59.879 laufenden Wortformen, das Flugblattkorpus 508 Texte mit 213.010 laufenden Wortformen (GerMov 2010). Die Suche nach *Maßnahme* in diesem Korpus zeigt im Kontrast zum obigen Kookkurrenzprofil klare Unterschiede: Die häufigsten linken Kollokatoren sind das Demonstrativpronomen *diese* und der bestimmte Artikel *die* sowie eine Reihe von Adjektiven wie *administrative, polizeiliche, drakonische* etc. Diese Kollokatoren zeigen, dass in den untersuchten Dokumenten Maßnahmen von Behörden angeprangert und kritisiert werden. Der Sprachgebrauch ist also ganz anders als im Deutschen Referenzkorpus und zeigt eine Facette des spezifischen Charakters des 68er-Diskurses.

[2] Der Log-Likelihood-Signifikanztest prüft, ob sich die beobachteten und statistisch erwartbaren Häufigkeiten signifikant voneinander unterscheiden. Je höher der Wert, desto signifikanter ist der Unterschied.

3 Beispielanalyse

Nach den methodisch-theoretischen Erläuterungen oben wollen wir anhand kurzer Beispiele, die jedoch im Kontext umfangreicherer Studien stehen, zeigen, wie Diskurse korpuslinguistisch analysiert werden können. Wir gliedern die Analysen nach drei Methodenbündeln: Schlagwörter, Sprachgebrauchsmuster und Visualisierung von Sprachdaten.

3.1 Schlagwörter

Schlagwort-Analysen sind eine einfache Methode, um das spezifische Vokabular eines Korpus zu eruieren. Grundlage ist die statistische Berechnung der Lexeme, die bezüglich ihrer Frequenz im Korpus signifikant häufiger auftreten als in einem zum Vergleich herangezogenen Referenzkorpus. Dabei ist nicht die absolute Frequenz ausschlaggebend, sondern der Frequenzunterschied in zwei Korpora in Relation zu den jeweiligen Korpusgrößen. Damit können auch sehr niedrigfrequente Lexeme Schlagwörter sein, sofern der Frequenzunterschied zum Referenzkorpus signifikant ist.

Als Beispiel für solche Analysen sei auf zwei Studien (Bubenhofer/Schröter 2010; Bubenhofer/Scharloth 2011) im „Text+Berg"-Korpus (Bubenhofer u. a. 2011) verwiesen: Die Studien untersuchten den Wandel des „Sprechens über Berge", also den alpinistischen Diskurs in der Schweiz auf der Basis eines Korpus des Periodikums des Schweizer Alpenclubs. Das Gesamtkorpus umfasst 196 Bände, was einer Menge von 35,8 Mio. laufenden Wortformen entspricht. Die Studien untersuchten unterschiedliche Perioden des gleichmäßig definierten Rasters von 20-Jahre-Schritten (1880-1899, 1900-1919 etc.), wobei hier Ergebnisse vierer Perioden zusammengefasst werden.

In Abbildung 1 sind die Schlagwörter abgebildet, die für die Zeit von 1880 bis 1899 typisch sind, wenn sie mit den Daten von 1930 bis 1949 verglichen werden. Die Schlagwörter für die jüngere Periode sind in Abbildung 2 abgebildet. Alle abgebildeten Lexeme sind jeweils hoch sigifikant für das Korpus (LLR, $p < 0{,}0001$), wobei die Größe den Grad der Sigifikanz ausdrückt (und nicht etwa die pure Frequenz).

Abbildung 1: Schlagwörter im Text+Berg-Korpus der Periode 1880 bis 1899 im Vergleich zur Periode 1930 bis 1949 (LLR, $p < 0{,}0001$)

Auf den ersten Blick sind orthographische Veränderungen sichtbar (*Theil, Section, Thäler*), die nicht weiter interessieren. Im Kontrast zur früheren Phase fallen jedoch bei

der Epoche 1930–1949 (Abbildung 2) auf den zweiten Blick eine Reihe von Lexemen aus dem Wortschatz des Seelenlebens und der Gefühle auf: *Seele, Erlebnis, Herz, Erleben, Freude, Gefühl, Sehnsucht, Liebe, Glück, Einsamkeit, Mut, Empfinden, Gemüt* etc. In Lexemen wie *Leben, Schönheit, Wunder, Geheimnis, Menschlein, Romantik, Erfüllung, Unendlichkeit, Gott* und *Glaube* scheint ein religiös-romantisches Verhältnis des Menschen zur Natur anzuklingen. Personenbezeichnungen wie *Kamerad, Freund, Bergkamerad* und *Gefährte* betonen die soziale Verbundenheit und Gemeinschaft der Alpinisten. *Ski, Haken, Lawine, Überhang, Seillänge* oder *Steigeisen* verweisen auf neue Fortbewegungsarten, Betätigungsfelder und Techniken bei Bergtouren (Bubenhofer/Schröter 2010: 272).

In der früheren Epoche verwendete Lexeme wie *Itinerar, Studie* bzw. *Studium, Erkundigung, Reisebericht, Zeitangabe, Annahme* und *Beobachtung* lassen andere, explorativere Formen des Alpinismus vermuten. Auffallend ist auch, dass *Führer* und *Clubist* in der zweiten Periode stark zurückgegangen sind und *Alpinist*, *Bergsteiger* und *Bergführer* zugenommen haben. Dies könnte auf veränderte Funktionen, Selbst- und Fremdbilder der Akteure zurückgeführt werden (Bubenhofer/Schröter 2010: 272).

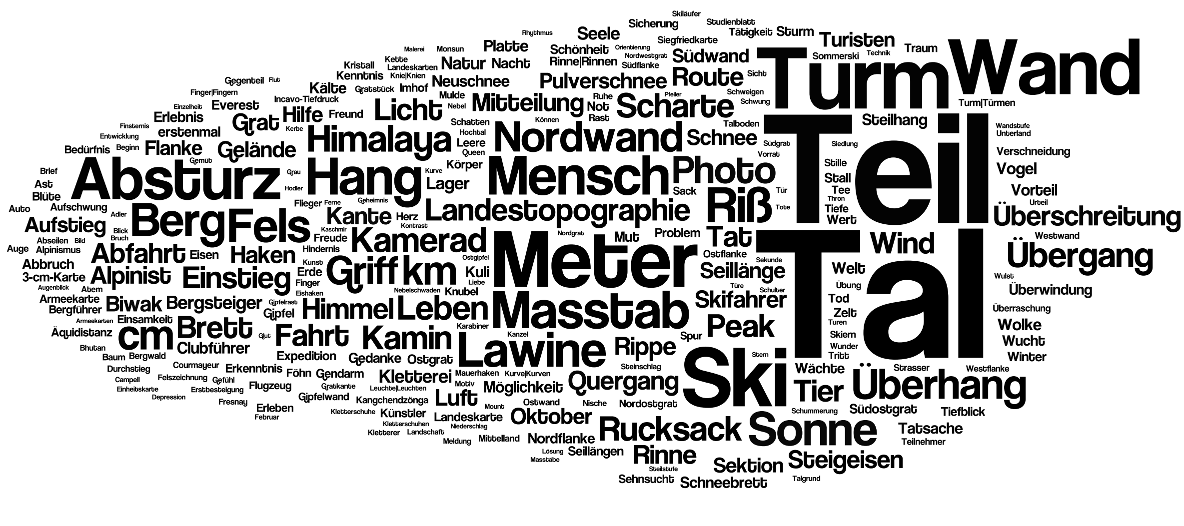

Abbildung 2: Schlagwörter im Text+Berg-Korpus der Periode 1930 bis 1949 im Vergleich zur Periode 1880 bis 1899 (LLR, p < 0,0001)

Solche Schlagwortwolken geben einen ersten Eindruck über die lexikalische Spezifik des Korpus. Weitere Analysen sind jedoch notwendig. Die Schlagwortlisten lassen sich beispielsweise dank eines vorgängigen Wortarten-Taggings der Korpusdaten nach Wortarten filtern. So zeigt sich z.B. in den 1930er/40er-Jahren generell eine höhere Nutzung von Personalpronomen, wobei die Zunahme der Frequenzen von *dein*, *du* und *wir* besonders signifikant ist (Bubenhofer/Schröter 2010: 272f.). Die Deutung dieser Befunde wird einfacher, wenn man sich vom Einzellexem löst und Sprachgebrauchsmuster in den Blick nimmt.

3.2 Sprachgebrauchsmuster

Unter „Sprachgebrauchsmustern“ verstehen wir Einheiten bestehend aus mehreren linguistischen Entitäten wie Wörtern, Wortarten, Lemmata, Tempus- oder Modusinformationen etc., die in einem Korpus musterhaft, also rekurrent, verwendet werden und pragmatisch gedeutet werden können (Bubenhofer 2009: 43 ff.). Wir operationalisieren diese Sprachgebrauchsmuster als „komplexe n-Gramme“, also Ketten von Wörtern: Komplexe n-Gramme sind eine erweiterte Form von n-Grammen, die nicht nur aus einer Folge von Wortformen bestehen, sondern auch aus einer Kombination von

Wortformen und Wortart-Informationen bestehen können. Während ein n-Gramm beispielsweise als Wortformenkette *so verbringen wir* definiert ist, werden bei der Berechnung von komplexen n-Grammen die Wortarten mit einbezogen, so dass eine Reihe von ähnlichen Wortformen-n-Grammen abstrakter als „*so* – finites Verb – Personalpronomen" gefasst werden.[3]

Für die Teilkorpora wurden alle komplexen n-Gramme berechnet, die jeweils für das Teilkorpus im Vergleich zu den anderen Teilkorpora typisch sind. Die Kategorisierung dieser n-Gramme führt in der Folge zu Sprachgebrauchsmustern, wenn sie korpuspragmatisch gedeutet werden können.

In den bereits erwähnten Studien zum Text+Berg-Korpus (Bubenhofer/Schröter 2010; Bubenhofer/Scharloth 2011) ergaben die Analysen der komplexen n-Gramme interessante Befunde: Zusammenfassend lässt sich sagen, dass für die Zeit von 1960 bis 1979 im Vergleich zur Periode 1990 bis 2009 Muster typisch sind, die für einen narrativen Stil stehen und dabei eine sehr persönliche Erzählperspektive wiedergeben. Denn die Muster weisen viele Personalpronomen der 1. Person Singular und Plural und Personennamen sowie typisches Wortmaterial für Erzählungen wie (meist: temporale und koordinierende) Konnektoren auf. Diese Beobachtung deckt sich mit den Spezifika des Teilkorpus mit den Texten der 1930er/40er-Jahre, in dem ebenfalls auf ähnliche Weise Geschichten erzählt werden (Bubenhofer/Schröter 2010).

Auffällig für das Teilkorpus 1960 bis 1979 ist z.B. das Muster „Adverb – finites Verb – irreflexives Personalpronomen", das im Korpus hochfrequent ist (5383 Mal) und im Teilkorpus hoch signifikant häufiger vorkommt als im Korpus 1990 bis 2009 (LLR 1667, $p < 0.00001$). Es steht für Realisierungen wie *so verbringen wir*, *so sehnten wir*, *dann stiegen wir* etc. Das Muster konzentriert sich zudem nicht auf wenige Autoren, sondern kommt bei 489 unterschiedlichen Autoren und in 758 unterschiedlichen Artikeln vor – gestreut über alle 20 Jahre des Teilkorpus. In den folgenden Tabellen sind eine Reihe von Mustern dieser Art mit Belegen abgedruckt (vgl. Bubenhofer/Scharloth 2011):

ADV VVFIN PPER[4]	Frequenzen absolut: 5383 (2688); LLR: 1666.761182; p < 0
	So verbringen wir trotz der scharfen Kälte eine letzte Nacht auf dem Gletscher auf 5000 m Höhe. (Band 1964, "Der Pico Unis West und der Tocllaraju", J.-J. Fatton) **So sehnten wir** die Stunde des Aufbruchs herbei, ahnten wir doch, dass uns einer der schönsten Anstiege unserer Alpen bevorstand und wir in einem idealen Crescendo dem grossartigen Gipfel der Grandes Jorasses zustreben würden. (Band 1977, "Zwei Tage auf 4000 Meter (Grandes Jorasses)", Philippe Staub) **Dann stiegen wir** mit der silbernen Kabine in lautloser Fahrt über die verträumten Weiler Winkelmatten und Blatten, zur Zwischenstation Furi, dann nach dem Umsteigen über Hermettje nach Schwarzsee an den Fuss des Matterhorns hinauf, das da so nahe steht, dass man ehrfürchtig-staunend es mit der Hand zu berühren glaubt. (Band 1962, "Über den Anderson- zum Lauteraargrat", Richard Knecht) **Damals umschwirrten uns** Fragen, Erwartungen, Hoffnungen ; (Band 1979, "Die Via Lepontina: eine Route in den Tessiner Alpen", Claudio Abächerli)
Unterschiedliche Autoren: 489; Artikel: 758; Jahre: 20	

[3] Zur Berechnung der komplexen n-Gramme wurde ein von der Forschergruppe *semtracks* entwickeltes Programm „cwb-n-grams" verwendet, das eine Erweiterung zur Open Corpus Workbench ist (Evert/The OCWB Development Team 2010).

[4] Vergleiche zu den Abkürzungen der Wortarten das STTS-Tagset (Schiller/Teufel/Stöckert/Thielen 1999), das auch übers Web einsehbar ist: http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html (27. Oktober 2011).

,/$, ADV dass/KOUS PPER	Frequenzen absolut: 439 (69); LLR: 421.061016; p < 0
	Ich war ziemlich müde vom Aufstieg, **so dass wir** bis zur Gianettihütte (oder Capanna Badile, 2534 m) fünf Stunden benötigten. (Band 1973, "Ein unbeabsichtigter Rekord", W. Kirstein) Das Terrain war wenig schwierig, **so dass ich** etwas unaufmerksam kletterte und einen aus dem Riss herausragenden grossen Stein berührte. (Band 1962, "Auf der Spannorthütte", W. Müller-Hill)

ADV VMFIN PPER	Frequenzen absolut: 840 (429); LLR: 246.851027; p < 0
	Zuerst mussten wir uns zwar auch hier über eine fast zwei Meter hohe Wächte hinunterschwindeln, dann stob der Schnee gegen das Gesicht, und wir sahen nichts mehr, bis wir sanft in einer Mulde landeten. (Band 1976, "Die Reise zum Gran Sasso d'Italia", Alfred Graber) Doch **inzwischen dürften sie** im Spital Visp in guter Obhut sein. (Band 1971, "Bergkameradschaft", Ernst Bucher) **Dort musste ich** einen Wagen nehmen, der mich in richtigem Fieberzustand nach Martigny brachte. (Band 1963, "Eine Bergsteigerkarriere vor 100 Jahren: G. Ad. Koella, 1822-1905", Louis Seylaz) **Heute darf ich** ja ruhig zugeben, dass es seine Hartnäckigkeit ist, die mir das Vertrauen in ihn geschenkt hat. (Band 1978, "In den Schluchten des Verdon. Paroi du Duc", Michel Pétermann) Sonntags ruhen wir uns in Lima aus, **dann müssen wir** die Vorbereitungsarbeiten abschliessen, damit wir die Reise nach Monterrey antreten können. (Band 1974, "Cordillera Bianca – bezaubernd und unvergesslich", E. Borioli) **Jetzt müssen wir** uns aber beeilen. (Band 1972, "Rosenlauistock-Südwestwand", Andreas Flückiger)
Unterschiedliche Autoren: 304; Artikel: 415; Jahre: 20	

./$. Als/KOUS PPER	Frequenzen absolut: 398 (198); LLR: 122.157474; p < 0
	Ich kam mir sehr schlau vor und fühlte mich meiner Sache sicher. **Als ich** am Samstagabend die Argentierehütte betrat, war ich beides nicht mehr, dafür aber nass, pudelnass wie noch nie in meinem Leben. (Band 1970, "Die Aiguille Verte", Pierre Vittoz) Der Wunsch, jene Gegenden kennenzulernen, wurde immer heftiger, und ich überprüfte meinen Ferienkalender. **Als ich** auch Carlos, meinen jungen venezolanischen Freund, der mich auf allen Abenteuern in Bolivars Landen begleitet hatte, für die Idee begeistern konnte, stand das Ziel fest : (Band 1967, "Auf Edward Whympers Spuren in Ecuador", Fritz Aeberli) Der nun aufgegangene Mond leuchtete uns, während wir über den Gletscher abstiegen. **Als wir** aber auf die riesige Moräne kamen, wurde sein Licht ungenügend, und mehr kugelnd und oft fallend, ohne aber den geringsten Schaden davonzutragen, erreichten wir eine Stunde vor Mitternacht das Basislager. (Band 1960, "Nevado Panta", Geny Steiger)
Unterschiedliche Autoren: 176; Artikel: 231; Jahre: 20	

Im Gegensatz dazu stehen die Muster des jüngeren Korpus von 1990 bis 2009: Hier herrscht ein deskriptiver Stil vor und es sind mehr Passiv-Konstruktionen zu beobachten als im älteren Korpus.

Bei Hypothesen wie jener über die Passiv-Konstruktionen ist es hilfreich, systematische Analysen vorzunehmen: So lässt sich mit einer maschinellen Analyse, die die automatische Wortarten-Annotation des TreeTaggers (Schmid 1994) weiter auswertet,

ein Passivkoeffizient ausrechnen. Wie folgende Grafik zeigt, ist tendenziell eine Abnahme der Passivkonstruktionen feststellbar, wobei ab den 1955er-Jahren ein Anstieg zu beobachten ist.

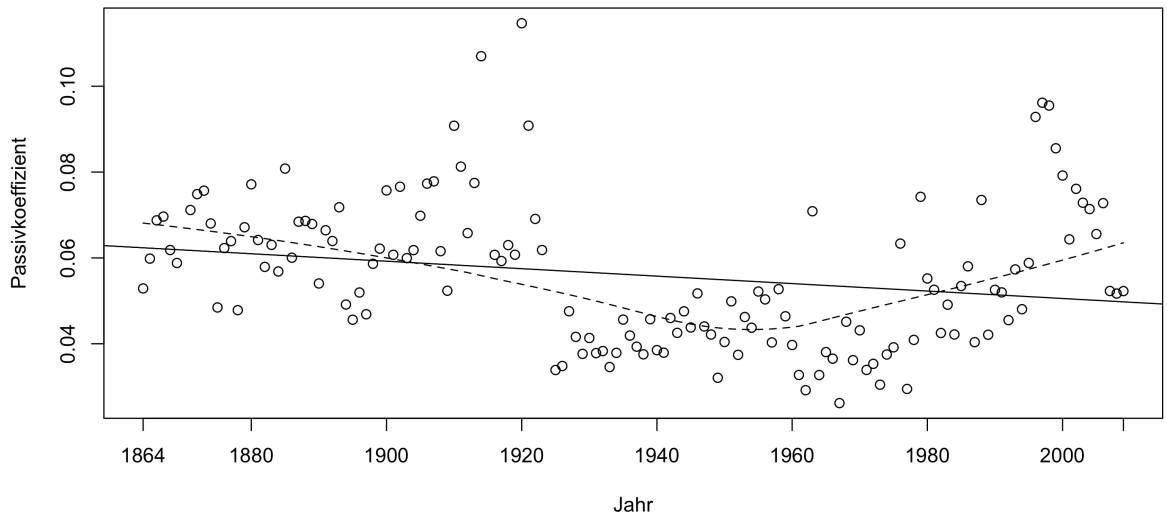

Abbildung 3: Durchschnittlicher Passivkoeffizient pro Jahr mit Trendlinien

Vergleicht man die beiden Zeitperioden zwischen 1960 bis 1979 und 1990 bis 2009 miteinander, ergibt sich ein hochsignifikanter Unterschied der Passivkoeffizienten. In den neuen Daten sind mehr Passivkonstruktionen vorhanden (t = -13.2002, df = 1997.302, p < 2.2e-16).

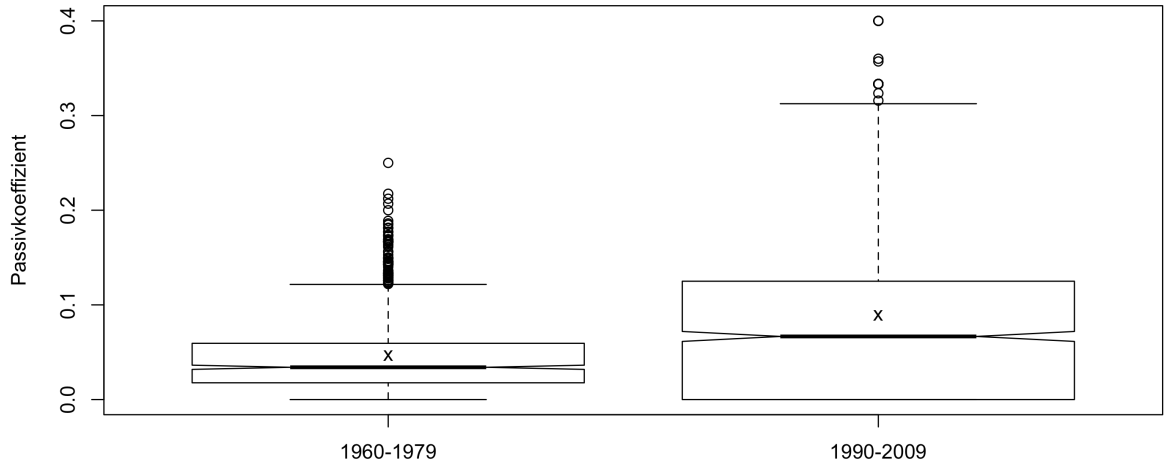

Abbildung 4: Durchschnittlicher Passivkoeffizient in zwei Perioden im Vergleich

In Kombination mit den Befunden der Schlagwortanalyse ergaben die komplexen n-Gramme und weitere hypothesengeleitete Untersuchungen (z.B. zur Verwendung von lexematischen Intensivierern[5] wie *absolut*, *komplett*, *höchst*, *sagenhaft*, *schrecklich* etc., der durchschnittlichen Satzlänge etc.) die folgenden Hypothesen über die Veränderung des Diskurses „Sprechen über Berge“:

[5] Vgl. Biedermann (1969), Bierwisch (1987), (Os) 1989.

1. „Bergsteigen-/wandern wird für SAC-Mitglieder von einer objektiven Erkundung ungewöhnlicher äußerer Gegebenheiten zunehmend zu einem subjektiven Erlebnis ungewohnter äußerer Umstände“ (Bubenhofer/Schröter 2010: 278). Während im ausgehenden 19. Jahrhundert der wissenschaftliche Entdeckergeist vorherrschte, mit dem in sachlicher und systematischer Weise die Bergwelt entdeckt wurde, steht ab den 1930er-Jahren und noch stärker ab den 1960ern der Mensch im Vordergrund: Es ist seine innere Welt, die in Reaktion zum äußeren Erleben dargestellt wird.[6]
2. „Bergsteigen/-wandern wird für SAC-Mitglieder von einer bildungsbürgerlich fundierten Freizeitwissenschaft zunehmend zu einem professionell betriebenen Extremsport“ (Bubenhofer/Schröter 2010: 278). In neuster Zeit ab den 1990er-Jahren zeigt sich ein utilitaristischer Zugang zu den Bergen, bei dem diese als Objekte der Freizeitgestaltung und des sportlichen Wettkampfs betrachtet werden, weshalb Tipps und Vorschläge gefragt sind, wie die Freizeit am besten gestaltet werden kann.

Diese Thesen mögen vielleicht nicht überraschend sein, doch lassen sie sich mit einer datengeleiteten, linguistischen Perspektive erarbeiten, und es ist mit Hilfe der Korpuslinguistik auch möglich aufzuzeigen, mit welchen sprachlichen Mitteln dieser Diskurs konstruiert wird.

3.3 Visualisierung

Eine wichtige Möglichkeit, einen Überblick beim Betrachten von quantitativen Analyseresultaten zu gewinnen, scheint uns das Mittel der Visualisierung zu sein. Sogenannte „bildgebende Verfahren“ sind in anderen Disziplinen als Mittel der Erkenntnisgewinnung weit verbreitet, so beispielsweise in der Medizin. In den Sozialwissenschaften gewinnen Techniken der Visualisierung z.B. im Bereich der Netzwerkanalysen, an Bedeutung (Brandes/Freeman/Wagner im Druck).

In den Kultur- und Geisteswissenschaften zeichnet sich ein riesiges Potenzial für bildgebende Verfahren ab, das noch kaum genutzt wird. Papenbrock und Scharloth (2011) zeigen anhand datengeleiteter Analysen von Ausstellungskatalogen zur NS-Zeit, wie Visualisierungen helfen, die kunsthistorischen Daten zu deuten. Durch eine Kollokationsanalyse, bei der das Kovorkommen von Künstlernamen auf Ausstellungskatalogen berechnet und anschließend in Graphen visualisiert wird, ist es möglich, die Vernetztheit und damit die Ausstellungslandschaft während der NS-Zeit sichtbar zu machen.

[6] Günther (1998: 13, 161), die sich in ihrer Studie des „bürgerlichen Alpinismus (1870-1930)“ stark auf die Schriften des Deutschen und Österreichischen Alpenvereins (DÖAV) stützt, spricht dementsprechend davon, dass „der Entdeckeralpinismus [um die Jahrhundertwende] vom Schwierigkeitsalpinismus abgelöst wird“.

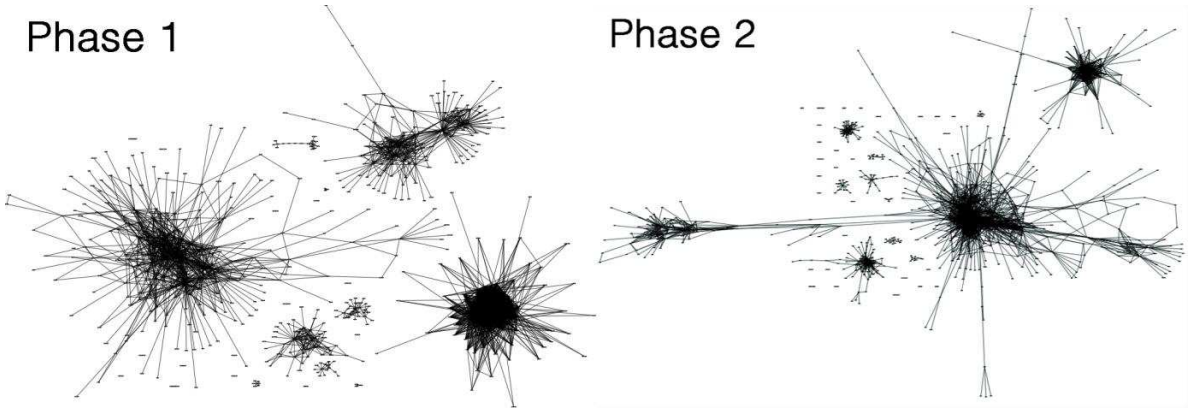

Abbildung 5: Die Vernetztheit von Künstlern in den Zeiträumen 1933 bis 1937 (Phase 1) und 1938 bis 1945 (Phase 2) berechnet auf der Basis von Ausstellungskatalogen (Papenbrock/Scharloth 2011, S. 36)

Wie Abbildung 5 zeigt, bietet das bildgebende Verfahren gegenüber der Auflistung aller Kollokatoren einen Mehrwert. Es lassen sich in den beiden Phasen zwei unterschiedliche Konfigurationen von Verdichtungen mit unterschiedlichen Verhältnissen von Zentrum und Peripherie ausmachen: In der ersten Phase ordnen sich die Verdichtungen an der Peripherie an und weisen kaum Verbindungen untereinander auf, das heißt, die regionale Orientierung des Ausstellungswesens war offenbar in den ersten Jahren des Nationalsozialismus ausgeprägter, das Ausstellungswesen eher dezentral als zentralistisch strukturiert. Das Kollokationsnetz zur Phase 2 zeigt dagegen eine starke Verdichtung, von der aus Verbindungen zu kleineren Verdichtungen an den Rändern bestehen. Hier zeigen sich die Effekte der auf Vereinheitlichung ausgerichteten nationalsozialistischen Kunst- und Ausstellungspolitik sehr deutlich. Die starke Verdichtung im Zentrum des Graphen kann als ein Indikator für eine Homogenisierung des Ausstellungswesens, für eine Auflösung der regionalen Strukturen und damit für eine tendenziell vereinheitlichende, d.h. nationale Entwicklung gelesen werden (vgl. dazu Papenbrock/Scharloth 2011, S. 36).

Es lässt sich erahnen, welches Potenzial solche Verfahren für diskurslinguistische Belange bietet. Natürlich muss man sich dabei nicht nur auf die korpuslinguistische Aufbereitung der Daten stützen (also auf linguistische Einheiten), sondern es können beliebige außersprachliche Faktoren zu Texten miteinbezogen werden, um z.B. über statistische Cluster- oder Faktorenanalysen die Charakteristika von Texten, Akteuren oder Institutionen zu berechnen.

Zwei weitere Beispiele von Visualisierungen sprachlicher Daten sollen zum Schluss noch herangezogen werden. Abbildung 6 zeigt Daten aus einer umfangreichen korpuslinguistischen Analyse eines Korpus von knapp 45.000 Artikeln der Neuen Zürcher Zeitung (27,9 Mio. laufende Wortformen) von 1995 bis 2005. Die Abbildung visualisiert die Nennung von Nationalitäten als Personenbezeichnungen im Auslands-Ressort (5670 Artikel, 3,4 Mio. laufende Wortformen) und stellt sie auf einer Karte dar (Bubenhofer 2009, S. 279). Genauer werden dabei nicht einfach die Frequenzen der Nennungen dargestellt, sondern das Ausmaß der Veränderung der Frequenzen im Korpus (Variationskoeffizient): Je größer der dargestellte Punkt auf der Karte, desto stärker variiert die Häufigkeit, mit der die entsprechende Nationalität im Korpus genannt wird.

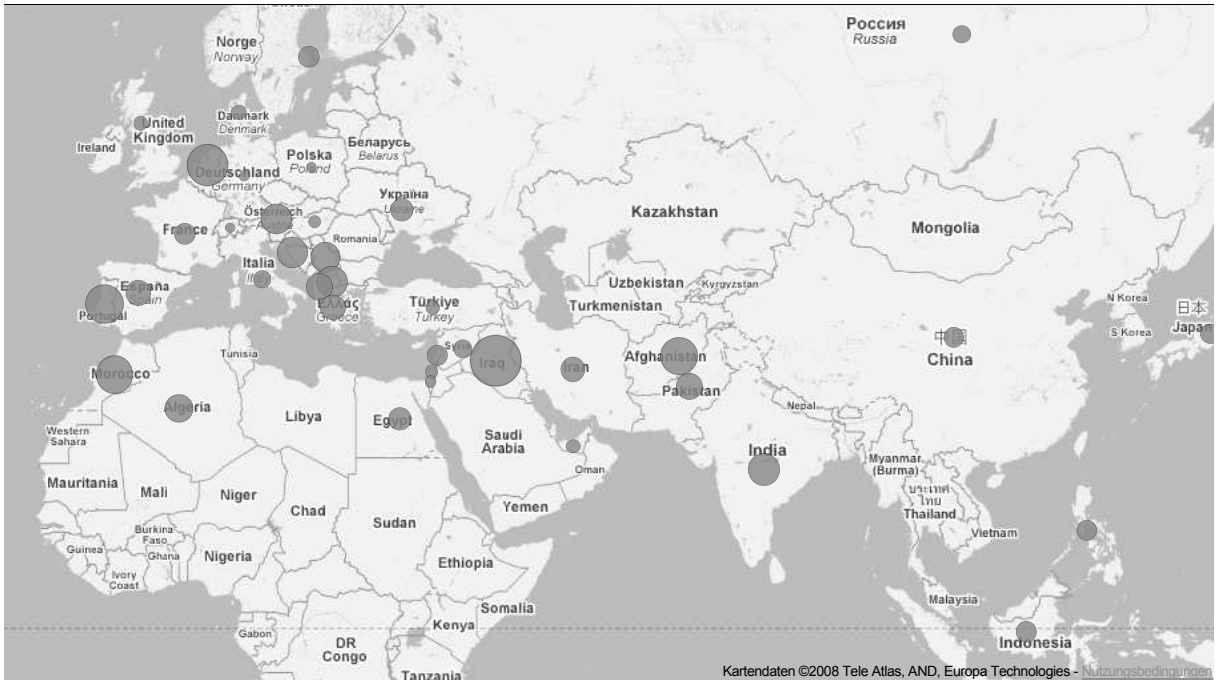

Abbildung 6: Neue Zürcher Zeitung, Ressort „Ausland“: Die größten Veränderungen der Nennungen von Einwohner/innen (Nationalitäten) im NZZ-Korpus (Variationskoeffizient $s/\bar{x}$, N = 3205, $\bar{x} \geq 1$).

Dabei ergeben sich sofort Hinweise für die Hypothese, dass Bezeichnungen für Nationalitäten oder Ethnien ein Indikator für kriegerische oder zumindest problematische Kontexte sind. Die Karte ist demnach ein Abbild unserer mentalen Landkarte der konfliktträchtigen Regionen.[7]

Im zweiten Beispiel in Abbildung 7 werden wiederum Daten aus dem Text+Berg-Korpus verwendet, in dessen Kontext bereits weiter oben Analysen präsentiert wurden. Als Datengrundlage dient ein mit geografischen Informationen ausgezeichnetes Korpus: Durch eine maschinelle Analyse wurden geografische Entitäten wie Berge, Gletscher und Hütten annotiert (Volk u. a. 2010) und mit geografischen Koordinaten versehen. So wurde es möglich, die Kollokatoren zu diesen Entitäten zu berechnen, um eine Karte zu zeichnen, die den Sprachgebrauch wiedergibt, mit dem diese Orte beschrieben werden.

[7] Die Basis für die Annotation der Nationalitätenbezeichnungen waren das Verzeichnis des amtlichen Sprachgebrauchs Deutschlands (Sprachendienst des Auswärtigen Amts 2006) und die „Liste der Hauptstädte der Welt“ der Wikipedia (Freie Enzyklopädie Wikipedia). Diese Daten wurden manuell ergänzt um inoffizielle, aber gebräuchliche Bezeichnungen wie *Libyen*, *USA* oder *Jugoslawien* etc. Vgl. Bubenhofer (2009: 195).

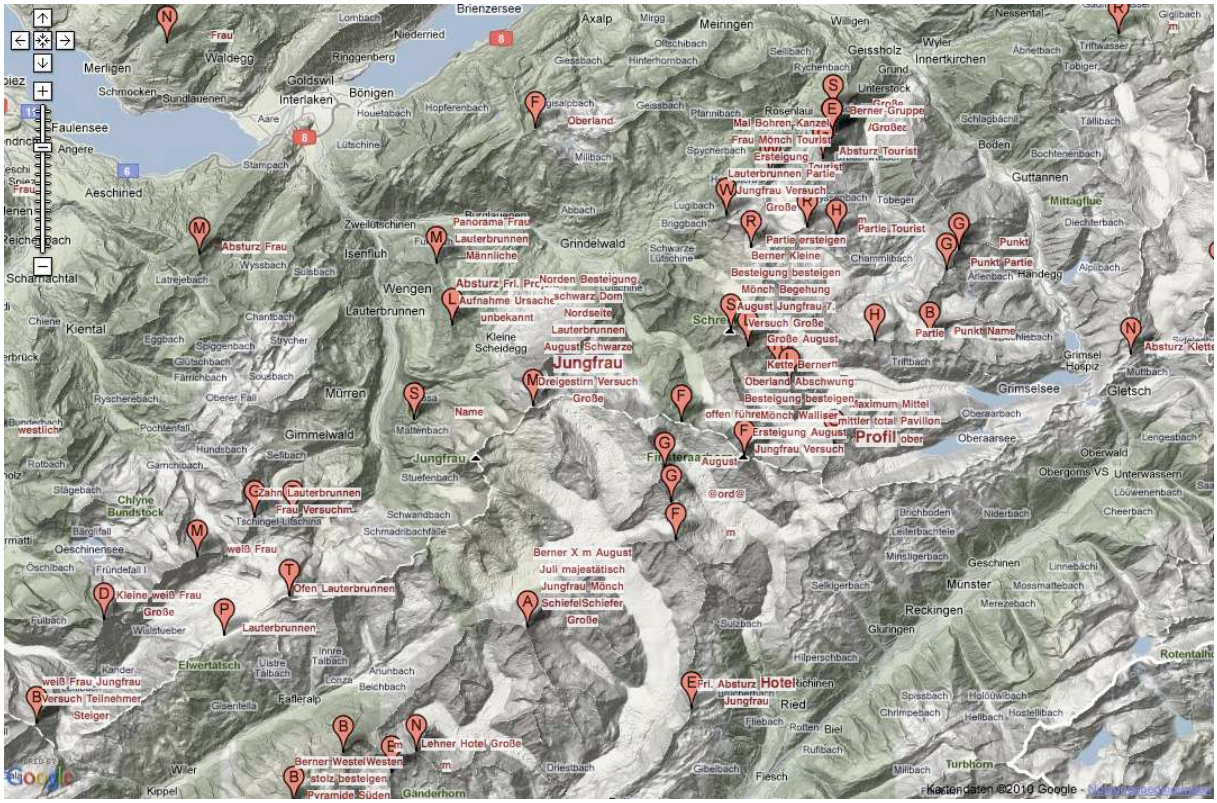

Abbildung 7: Kollokatoren zu Bergen visualisiert auf einer Karte

Da die Karte in digitaler Form besteht, können die Daten zudem dynamisch und interaktiv dargestellt werden. Die Darstellung kann beeinflusst werden, indem z.B. die Kollokatoren nach Wortarten gefiltert oder der statistische Schwellenwert, ab dem Kollokatoren dargestellt werden sollen, beeinflusst wird. Als Mehrwert gibt die Karte auch einen Eindruck davon, über welche Regionen besonders oft oder selten berichtet wird, wobei, dank der geografischen Referenzierung, ein geografischer Zusammenhang zwischen den Daten ersichtlich wird, der über die Durchsicht von reinen Textlisten von Kollokatoren nicht in der gleichen Weise deutlich wäre. So visualisiert die Karte geografische Cluster von ähnlichen Kollokatoren, während bei der Durchsicht von textbasierten Listen die geografische Nähe der Kollokatoren verborgen bliebe.

4 Fazit

Spitzmüller und Warnke plädieren dafür, dass in der Diskurslinguistik quantitative Verfahren die qualitativen Analysen nicht ersetzen, sondern ergänzen sollen (Spitzmüller/Warnke 2011: 39). Wir glauben, dass der Nutzen quantitativer Analysen oft stark unterschätzt wird und es lohnenswert ist, den vielfältigen Methodenapparat aus der Korpuslinguistik, der Statistik (Data Mining-Verfahren, Clusteranalysen, Faktorenanalysen etc.) sowie der Datenvisualisierung für diskurslinguistische Fragestellungen fruchtbar zu machen. Es ist zu vermuten, dass sich nicht alle Forscherinnen und Forscher darauf einlassen wollen und es zudem zu einer Ausdifferenzierung innerhalb der Diskurslinguistik kommt: Einerseits in eine datenintensive quantitative Diskursanalyse, die die Ergebnisse statistischer Verfahren ernst nimmt und nicht nur als Inspirationsquelle für die Hypothesenbildung betrachtet (Scharloth/Eugster/Bubenhofer im Druck), andererseits in eine qualitative, die auf die Lektüre und präzise Analyse von Texten setzt.

Ein dritter Weg könnte die „quantitativ informierte qualitative Analyse“ sein, die darauf beruht, den menschlichen analytischen Leseprozess mit quantitativen Analysen zu unterstützen (Bubenhofer im Druck). Diese Idee ist nicht neu (vgl. z.B. Mautner 2012: 97) und scheint ein Desiderat zu sein: So argumentiert Mautner (2012: 90) im Kontext der kritischen Diskursanalyse (CDA), dass dem Vorwurf der ideologischen Vorgefasstheit begegnet werden könne, wenn die Analysen u.a. verstärkt quantitativ arbeiten, ohne den Nutzen qualitativer Analysen aufzugeben. Storjohann und Schröter (2011: 32) zeigen mit ihren diskurslinguistischen Arbeiten ebenfalls, dass quantitative mit qualitativen Methoden verknüpft werden können.

Gerade vor dem Hintergrund der Frage, wie deskriptiv oder kritisch eine Diskurslinguistik sein darf, ist ein methodisch sauberes Fundament wichtig. Wie wir zeigten, bietet ein quantitativer korpuslinguistischer Zugang die Chance, Daten auch induktiv zu analysieren und sich mit möglichst offenen Hypothesen dem Untersuchungsgegenstand zu nähern. Datengeleitete Analysen dieser Art können zu neuen Hypothesen führen, die vorher nicht im Bewusstsein des Forschers oder der Forscherin waren.

Noch wichtiger ist aber die Möglichkeit, große Datenmengen mit avancierten statistischen, korpus- und computerlinguistischen Methoden analysieren zu können. Die daraus resultierenden Ergebnisse und neuen Hypothesen fußen auf ausreichend vielen Beobachtungen, und die Methode, wie sie erreicht worden sind, ist transparent und reproduzierbar (vgl. auch Scharloth/Eugster/Bubenhofer im Druck).

Die *Deutung* der Ergebnisse kann anschließend sowohl eher dem deskriptiven als auch dem kritischen Paradigma der Diskurslinguistik folgen. Die quantitative Fundierung würde aber helfen, die Spannung zwischen den beiden Paradigmen abzubauen.

5 Bibliographie

Belica, Cyril (2001): Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. http://corpora.ids-mannheim.de/ccdb/ (4. September 2012).

Belica, Cyril/Steyer, Kathrin (2006): Korpusanalytische Zugänge zu sprachlichem Usus. In: AUC (Acta Universitatis Carolinae), Germanistica Pragensia XX.

Biedermann, Reinhard (1969): Die deutschen Gradadverbien. Diss. Heidelberg: Universität Heidelberg.

Bierwisch, Manfred (1987): Semantik der Graduierung. Grammatische und konzeptuelle Aspekte von Dimensionsadjektiven. In: Bierwisch, Manfred/Lang, Ewald (Hrsg.): Grammatische und konzeptuelle Aspekte von Dimensionsadjektiven. Berlin: Akademie-Verlag, 91–286.

Brandes, Ulrik/Freeman, Linton C./Wagner, Dorothea (im Druck): Social Networks. In: Tamassia, Roberto (Hrsg.): Handbook of Graph Drawing and Visualization. London. http://www.cs.brown.edu/~rt/gdhandbook/ (4. September 2012).

Bubenhofer, Noah (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. (Sprache und Wissen 4). Berlin, New York.

Bubenhofer, Noah (im Druck): Quantitativ informierte qualitative Diskursanalyse.

Korpuslinguistische Zugänge zu Einzeltexten und Serien. In: Roth, Kersten Sven/Spiegel, Carmen (Hrsg.): Perspektiven einer angewandten Diskurslinguistik. Berlin: Akademie-Verlag.

Bubenhofer, Noah/Dussa, Tobias/Ebling, Sarah (2009): „So etwas wie eine Botschaft." Korpuslinguistische Analysen der Bundestagswahl 2009. In: Sprachreport 4, S. 2–10.

Bubenhofer, Noah/Scharloth, Joachim (2011): Korpuspragmatische Analysen alpinistischer Literatur. In: Elmiger, Daniel/Kamber, Alain (Hrsg.): La linguistique de corpus – de l'analyse quantitative à l'interpretation qualitative / Korpuslinguistik – von der quantitativen Analyse zur qualitativen Interpretatino, Travaux neuchâtelois de linguistique 55, Neuchâtel: Institut des sciences du langage et de la communication, S. 241–259.

Bubenhofer, Noah/Schröter, Juliane (2012): „Die Alpen. Sprachgebrauchsgeschichte – Korpuslinguistik – Kulturanalye". In: Maitz, Péter (Hrsg.): Historische Sprachwissenschaft. Erkenntnisinteressen, Grundlagenprobleme, Desiderate. Studia Linguistica Germanica 110, Berlin/Boston: de Gruyter, S. 263–287.

Bubenhofer, Noah/Volk, Martin/Althaus, Adrian/Jitca, Magdalena/Bangerter, Maya/Sennrich, Rico (Hrsg.): Text+Berg-Korpus (Release 145). Digitale Edition des Jahrbuch des SAC 1864-1923 und Die Alpen 1925-2009. Institut für Computerlinguistik, Universität Zürich, 2011.

Evert, Stefan/The OCWB Development Team (2010): The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial. http://cwb.sourceforge.net/documentation.php http://cwb.sourceforge.net/documentation.php (4. September 2012).

Feilke, Helmuth (2000): Die pragmatische Wende in der Textlinguistik. In: Brinker, Klaus (Hrsg.): Text- und Gesprächslinguistik/Linguistics of Text and Conversation. (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science). Berlin/New York, S. 64–82.

Feilke, Helmuth/Linke, Angelika (Hrsg.) (2009): Oberfläche und Performanz. Untersuchungen zur Sprache als dynamische Gestalt. Berlin, New York.

Freie Enyklopädie Wikipedia: Liste der Hauptstädte der Erde. http://de.wikipedia.org/wiki/Liste_der_Hauptstädte_der_Erde (4. September 2012).

GerMov (2010): Ein Korpus zur gesprochenen und geschriebenen Sprache der 1968er-Bewegung in der Bundesrepublik Deutschland: Flugblätter und Tonbandprotokolle. Zusammengestellt und herausgegeben von Joachim Scharloth und Noah Bubenhofer. (= cosmov, Korpora zur Sprache sozialer Bewegungen 2). http://semtracks.com/korpora/ (4. September 2012).

Günther, Dagmar (1998): Alpine Quergänge: Kulturgeschichte des bürgerlichen Alpinismus (1870 -1930). Frankfurt am Main: Campus.

Institut für Deutsche Sprache (2010): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2010-I (Release vom 2.3.2010). Mannheim. http://www.ids-mannheim.de/kl/projekte/korpora/ (4. September 2012).

Lemnitzer, Lothar/Zinsmeister, Heike (2006): Korpuslinguistik. Eine Einführung. Tübingen: Narr.

Manning, Christopher D./Schütze, Hinrich (2002): Foundations of Statistical Natural Language Processing. 5th Ed. Cambridge, Massachusetts.

Mautner, Gerlinde (2012): Die kritische Masse. Korpuslinguistik und kritische Diskursanalyse. In: Felder, Ekkehard; Müller, Marcus; Vogel, Friedemann (Hrsg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen. Berlin, New York: de Gruyter (Linguistik – Impulse und Tendenzen).

Os, Charles van (1989): Aspekte der Intensivierung im Deutschen. Tübingen: Narr.

Papenbrock, Martin/Scharloth, Joachim (2011): Datengeleitete Analyse kunsthistorischer Daten am Beispiel von Ausstellungskatalogen aus der NS-Zeit: Musteridentifizierung und Visualisierung. In: Kunstgeschichte. Open Peer Reviewed Journal. http://www.kunstgeschichte-ejournal.net/248/ (4. September 2012).

Perkuhn, Rainer/Belica, Cyril (2006): Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik. In: Sprachreport 22 (1), S. 2–8.

Scharloth, Joachim (2011): 1968 - Eine Kommunikationsgeschichte. Paderborn: Fink.

Scharloth, Joachim/Bubenhofer, Noah (2012): Datengeleitete Korpuspragmatik: Korpusvergleich als Methode der Stilanalyse. In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hgg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin, New York: de Gruyter, S. 195-230.

Scharloth, Joachim/Eugster, David/Bubenhofer, Noah (im Druck): Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In: Busse, Dietrich/Teubert, Wolfgang (Hrsg.) (2013): Linguistische Diskursanalyse – Neue Perspektiven. Wiesbaden: VS-Verlag.

Storjohann, Petra; Schröter, Melani (2011): Die Ordnung des öffentlichen Diskurses der Wirtschaftskrise und die (Un-) Ordnung des Ausgeblendeten. In: Aptum. Zeitschrift für Sprachkritik und Sprachkultur. 7 (1), S. 32-53.

Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Stuttgart. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (4. September 2012).

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Spitzmüller, Jürgen/Warnke, Ingo H. (2011): Diskurslinguistik: eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. Berlin, New York: De Gruyter.

Sprachendienst des Auswärtigen Amtes (2006): Verzeichnis der Staatennamen für den amtlichen Gebrauch in der Bundesrepublik Deutschland. Stand: 18. Oktober 2006. Elektronische Ressource, Berlin, http://www.auswaertiges-amt.de/DE/Infoservice/Terminologie/Uebersicht_node.html (4. September 2012).

Steyer, Kathrin (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer, Kathrin (Hrsg.): Wortverbindungen – mehr oder weniger fest. (Institut für Deutsche Sprache. Jahrbuch 2003). Berlin, New York: De Gruyter, S. 87–116.

Tognini-Bonelli, Elena (2001): Corpus Linguistics at Work. (Studies in Corpus linguistics). Amsterdam: Benjamins.

Volk, Martin/Bubenhofer, Noah/Althaus, Adrian/Bangerter, Maya/Furrer, Lenz/Ruef, Beni (2010): Challenges in Building a Multilingual Alpine Heritage Corpus. In: Seventh

International Conference on Language Resources and Evaluation (LREC), Malta, 19 May 2010 - 21 May 2010, p. 1653-1659. https://www.zora.uzh.ch/34264/ (4. September 2012).